Genomic Data Lakehouse Architecture

Marek Wiewiórka, Agnieszka Szmurło, Tomasz Gambin

February 27, 2023

Not an Agenda *per se*…

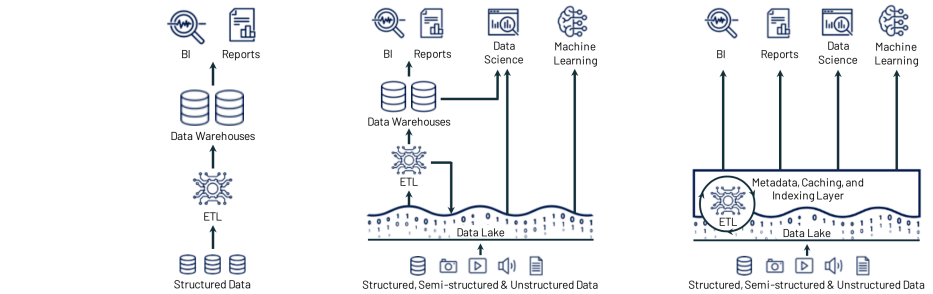
1. Does *Data Lakehouse* = Data Warehouse + Data Lake ?
2. Why does Genomic Data + Data Lakehouse ≠ *Genomic Data Lakehouse* ?
3. Is the SeQuiLa project to the rescue?

About me

- Chief Data Architect @GetInData | Part of Xebia
- Research Assistant[a] at Warsaw University of Technology
- putting the finishing touches to his Phd dissertation…
- keen long distance runner, gravel bikes enthusiast and absolutely in love with the Italian Lakes!

[a]Institute of Computer Science

Data platforms architectures



(a) First-generation platforms. **(b) Current two-tier architectures.** **(c) Lakehouse platforms.**

Figure: Evolution of data platform architectures to today's two-tier model (a-b) and the new Lakehouse model (c)[1]

Data Lakehouse design principles

- **low-cost** and **efficient** storage for very large-scale heterogeneous data backed by cloud object storage systems
- first-class support for **machine learning** and **data science** workloads with distributed DataFrame-like APIs
- state-of-the-art SQL **analytical** queries performance
- open **direct-access** file formats such as Parquet, ORC or DeltaLake

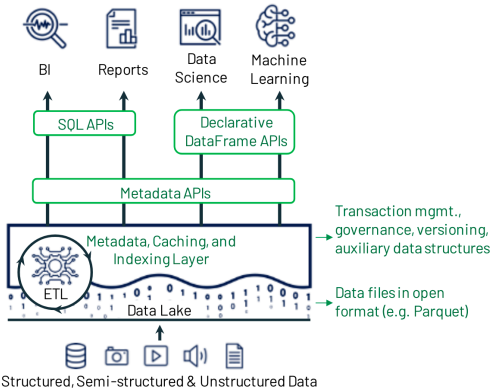


Figure: Data Lakehouse design[1]

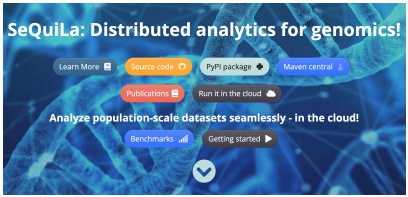
Challenges of Cloud Genomic Data Lakehouse

- genomic analyses primitives like e.g. **range** joins, depth of **coverage** or **pileup** summary
- genomic-specific file **formats** like BAM/CRAM or VCF
- secondary analyses pipelines - efficient **reusing** of the existing tools or **native** implementations
- GWAS-specific statistical methods
- set unified of unified APIs (**SQL** and **DataFrame**) – shift from traditional shell scripting/CLIs
- **cloud** challenges: legal, skills
- *ephemeral* computing vs long-lived infrastructure - the need of **IaC**[a] approach

[a] https://en.wikipedia.org/wiki/Infrastructure_as_code

The SeQuiLa project (1/2)

- originally started as a Phd project in 2018[a]
- meant for the elements of **secondary** and **tertiary** analysis
- extensions to the Apache Spark engine[8, 7, 6]
- **SQL** and **DataFrame** APIs
- support for bioinformatics file formats
- a set of **cloud** recipes[b]



[a] ://biodatageeks.github.io/sequila/
[b] https://github.com/biodatageeks/sequila-cloud-recipes

The SeQuiLa project (2/2)

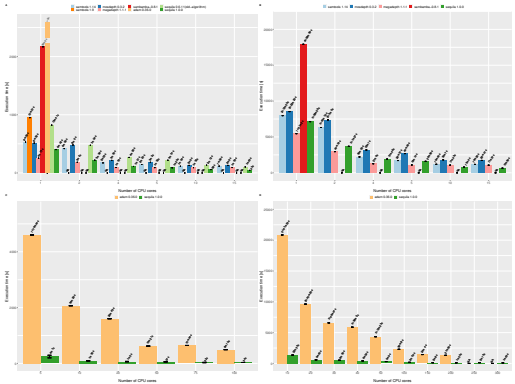


Figure: Depth of coverage[6]

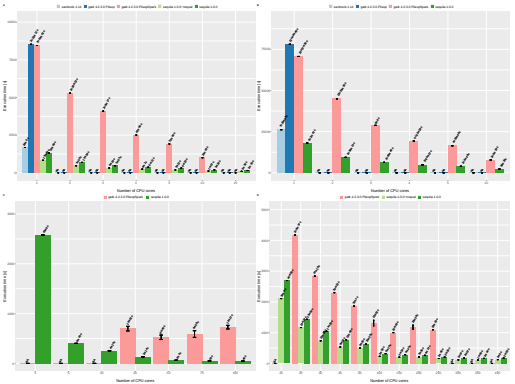
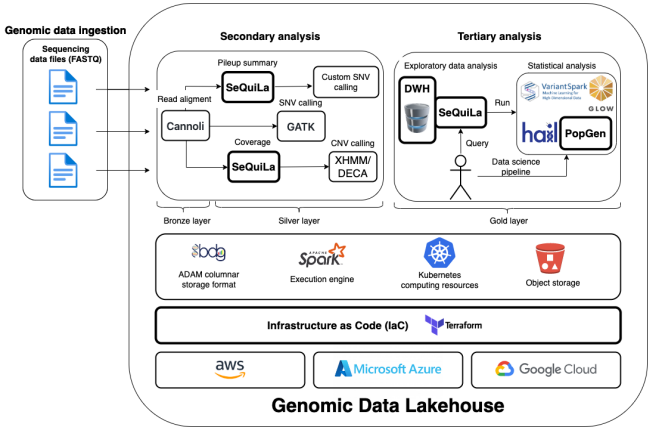


Figure: Pileup summary[6]

Cloud Genomic Data Lakehouse

- cloud-agnostic
- secondary and tertiary analysis
- *medallion architecture*[3]
- Apache Spark-based projects like ADAM[5], DECA[4], VariantSpark[2]



The 3 take-home messages

- Does *Data Lakehouse* = Data Warehouse + Data Lake ? – No[a]
- Why does Genomic Data + Data Lakehouse ≠ *Genomic Data Lakehouse* ? – It's complicated[b]
- Is the SeQuiLa project to the rescue? – Absolutely yes![c]

[a] a novel data platform architecture
[b] bioinformatics file formats and distributed genomic operations
[c] well, still some room for improvement

Thank you !

Q&A

marek.wiewiorka@gmail.com

Bibliography

[1] Michael Armbrust **andothers**. “Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics”. **in***Proceedings of CIDR*: 2021.

[2] Arash Bayat **andothers**. “VariantSpark: Cloud-based machine learning for association study of complex phenotype and large-scale genomic data”. **in***GigaScience*: 9.8 (**august** 2020). ISSN: 2047-217X. DOI: 10.1093/gigascience/giaa077.

[3] Ron L’Esteve. “Databricks”. en. **in***The Azure Data Lakehouse Toolkit*: Berkeley, CA: Apress, 2022, **pages** 83–139. ISBN: 978-1-4842-8232-8 978-1-4842-8233-5. DOI: 10.1007/978-1-4842-8233-5_3. URL: https://link.springer.com/10.1007/978-1-4842-8233-5_3 (**urlseen** 11/02/2023).

[4] Michael D. Linderman **andothers**. “DECA: Scalable XHMM exome copy-number variant calling with ADAM and Apache Spark”. **in***BMC Bioinformatics*: 20.1 (**october** 2019). Publisher: BioMed Central Ltd., **pages** 1–8. ISSN: 14712105. DOI: 10.1186/S12859-019-3108-7/TABLES/2. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3108-7 (**urlseen** 03/01/2023).

[5] Matt Massie **andothers**. “Adam: Genomics formats and processing patterns for cloud scale computing”. **in***University of California, Berkeley Technical Report, No. UCB/EECS-2013*: 207 (2013), **page** 2013.

[6] Marek Wiewiórka **andothers**. “Cloud-native distributed genomic pileup operations”. **in***Bioinformatics*: (**december** 2022). **byeditor**Peter Robinson. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac804. URL: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btac804/6900922.

[7] Marek Wiewiórka **andothers**. “SeQuiLa-cov: A fast and scalable library for depth of coverage calculations”. **in***GigaScience*: 8.8 (**august** 2019). ISSN: 2047-217X. DOI: 10.1093/gigascience/giz094. URL: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giz094/5543653.

[8] Marek Wiewiórka **andothers**. “SeQuiLa: an elastic, fast and scalable SQL-oriented solution for processing and querying genomic intervals”. **in***Bioinformatics*: 35.12 (**june** 2019). **byeditor**John Hancock, **pages** 2156–2158. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty940. URL: https://academic.oup.com/bioinformatics/article/35/12/2156/5182295.